République Algérienne Démocratique et Populaire
Ministère de l'enseignement supérieur et de la recherche scientifique

N° Attribué par la bibliothèque

Année univ.: 2018/2019

La régression en composantes principales et la régression PLS

Mémoire présenté en vue de l'obtention du diplôme de

Master Académique

Université de Saida - Dr Moulay Tahar

Discipline : MATHEMATIQUES

Spécialité : Analyse Stochastiques, Statistique des Processus et Applications (ASSPA)

par

Kabar **A**imaddine[1]

Sous la direction de

Dr/Mme F.Benziadi

Soutenu le 17/07/2019 devant le jury composé de

A. Bouaka	Université Dr Tahar Moulay - Saïda	Présidente
F. Benziadi	Université Dr Tahar Moulay - Saïda	Encadreur
S. Rahmani	Université Dr Tahar Moulay - Saïda	Examinatrice
R. Rouane	Université Dr Tahar Moulay - Saïda	Examinatrice

1. e-mail : kabarimad94@gmail.com

REMERCIEMENTS

Avant tous, je remercie Allah le tout-puissant qui a guidé tout au long de ma vie, qui a permis de m'instruire et d'arriver aussi loin dans les études, qui m'a donné courage et patience pour traverser tous les moments difficiles, et qui m'a permis d'achever ce travail.

Je tiens à exprimer toute ma reconnaissance à mon encadreur Madame *Benziadi Fatima*. Je la remercie de m'avoir encadré, orienté, aidé et surtout pour sa patience, sa disponibilité et ses judicieux conseils, qui ont contribué à alimenter ma réflexion.

Je remercie Docteur *Bouaka Aicha* pour l'attention qu'elle a manifestée à l'égard de ce mémoire, en s'engageant à en être la présidente de jury.

De même, je suis grandement reconnaissant à Docteur *Rahmani Saâdia* et Docteur *Rouane Rachida* de l'intérêt et du temps qu'elles ont bien voulu accorder à l'expertise de ce mémoire, en acceptant d'en être les examinatrices.

Je remercie mes très chers parents, qui ont toujours été là pour moi. Je remercie mes sœurs Amel, Hadjar et Nour Elhouda, et mon frère Larbi, pour leurs encouragements.

J'adresse mes sincères remerciements à tous les professeurs, intervenants et toutes les personnes qui par leurs paroles, leurs écrits, leurs conseils et leurs critiques ont guidé mes réflexions et ont accepté de me rencontrer et de répondre à mes questions durant mes recherches.

Enfin, je remercie mes amis Djillali, Merzoug, Nasredin, Cheikh, Abdouarahman, kheiredin, Toufik, Chamsedine, Housame, Zaine, qui ont toujours été là pour moi. Leur soutien inconditionnel et leurs encouragements ont été d'une grande aide.

À tous ceux qui ont, de prés ou de loin, participé à l'élaboration de ce mémoire, et que je ne pourrai nommer ici, je vous remercie de votre sollicitude aussi minime qu'elle ait pu être.

À tous, merci !

Table des matières

Introduction générale

En statistiques, en économétrie et en apprentissage automatique, un modèle de régression linéaire est un modèle de régression qui cherche à établir une relation linéaire entre une variable, dite expliquée, et une ou plusieurs variables, dites explicatives.

Parmi les modèles de régression linéaire les plus connus, le modèle de régression linéaire simple et le modèle de régression linéaire multiple.

En général, le modèle de régression linéaire désigne un modèle dans lequel l'espérance conditionnelle de Y sachant $X = x$ est une transformation affine en les paramètres.

Le modèle de régression linéaire est souvent estimé par la méthode des moindres carrés mais il existe aussi de nombreuses autres méthodes pour estimer ce modèle. On peut par exemple estimer le modèle par maximum de vraisemblance ou encore par inférence bayésienne.

Roger Joseph Boscovich(1755-1757)[10] est le premier scientifique a calculé les coefficients de régression linéaire, quand il entreprit de mesurer la longueur de cinq méridiens terrestres en minimisant la somme des valeurs absolues. **Pierre-Simon de Laplace**(1789)[10] utilise cette méthode pour mesurer les méridiens dans *Sur les degrés mesurés des méridiens et sur les longueurs observées sur pendule.*

La première utilisation de la méthode *des moindres carrés* est attribuée à **Adrien-Marie Legendre** en 1805 [21] ou à **Carl Friedrich Gauss**(1795) [10] qui dit l'avoir utilisée à partir de 1795.

Carl Friedrich Gauss(1821)[10] démontre le théorème connu aujourd'hui sous le nom de théorème de Gauss-Markov qui exprime sous certaines conditions la qualité des estimateurs, **Andrei Markov** le redécouvre en 1900.

L'origine du mot régression vient de Sir **Francis Galton**(1885)[13] dans son article *Galton exprime la taille des fils en fonction de la taille des pères.*

Plus tard la colinéarité des variables explicatives est devenue un sujet de recherche important. **Herman Wold**(1966)[30] propose un algorithme nommé tout d'abord NILES (Nonlinear estimation by Iterative Least Squares), puis NIPALS (Nonlinear estimation by Iterative Partial Least Squares), une des méthodes d'estimation conçues pour pallier la présence de colinéarité de certaines variables explicatives en imposant des contraintes sur les coefficients.

Arthur E. Hoerl et **Robert W. Kennard**(1970)[24] proposent la régression pseudo-orthogonale (Ridge Regression).

Herman Wold(1975)[30] présente l'approche PLS pour analyser les données exprimées en J blocs de variables sur les mêmes individus.

Svante Wold (fils d'**Herman Wold**) et **Harald Martens**(1983)[26] combinent NIPALS et l'approche PLS pour les adapter à la régression dans le cas où le nombre de variables est très supérieur au nombre d'observations (et où une forte multicollinearité est observée).

Svante Wold, **Nouna Kettaneh-Wold** et **Bert Skagerberg** (1989)[26] présentèrent pour la première fois la régression PLS non linéaire.

M. Stone et **R. J. Brooks**(1990)[26] proposent une méthode paramétrique permettant d'employer la méthode PLS pour la régression linéaire multiple, la PLS et la régression sur composantes principales.

La méthode du lasso (Lasso Regression), ayant le même objectif en utilisant une technique analogue, a été créée par **Robert Tibshirani**(1996)[29].

Les algorithmeses des méthodes de régression sur composantes (régression des moindres carrés partiels (PLS) et régression sur composantes principales (PCR)), recherchent des variables explicatives indépendantes liées aux variables initiales, puis estiment les coefficients de régression sur les nouvelles variables.

Dans ce travail, on parle du modèle de la régression linéaire multiple et le problème de la multicolinéarité. On propose deux méthodes d'estimation conçues pour pallier la présence de colinéarités de certaines variables explicatives.

Ce mémoire est organisé en quatre chapitres :

Dans le premier chapitre, on étudie le modèle linéaire multiple et l'estimation des paramètres de ce modèle par la méthode des moindres carrés pour des données linéairement indépendantes, et on présente le problème de la multicolinéarité.

Dans le deuxième chapitre, on donne un petit rappel sur l'analyse en composantes principales, et on parle de la première méthode de régression pour pallier le problème de multicolinéarité c'est la régression sur composantes principales (RCP).

Dans le même contexte, le troisième chapiter est consacré à létude de la deuxième solution qui est la régression aux moindres carrés partiels (PLS).

Une simulation pour comparer enter les trois méthodes de régression(MLR, RCP et PLS) à partir de l'erreur quadratique moyenne de prévision (MSEP) de chaque méthode sera présenté dans le quatrième chapiter.

Chapitre 1

Régression linéaire multiple (MLR)

1.1 Régression multiple

Le modèle de régression linéaire multiple est l'outil statistique le plus habituellement mis en oeuvre pour l'étude de données multidimensionnelles. Cas particulier de modèle linéaire, il constitue la généralisation naturelle de la régression simple.

1.1.1 Modèle

Une variable quantitative Y dite à *expliquer* (ou encore, réponse, exogène, dépendante) est mise en relation avec p variables quantitatives $X^1, ..., X^p$ dites *explicatives* (ou encore de contrôle, endogènes, indépendantes, régresseurs).

Les données sont supposées provenir de l'observation d'un échantillon statistique de taille n $(n > p + 1)$ de $\mathbb{R}^{p+1}$:

$$(x_i^1, ..., x_i^j, ..., x_i^p, y_i) \qquad i = 1, ..., n.$$

L'écriture du *modèle linéaire* dans cette situation conduit à supposer que l'espérance de Y appartient à un sous-espace de $\mathbb{R}^n$ engendré par $\{\mathbb{1}, X^1, ..., X^p\}$ où $\mathbb{1}$ désigne le vecteur de $\mathbb{R}^n$ constitué de "1". C'est-à-dire que les $(p + 1)$ variables aléatoires vérifient :

$$y_i = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + ... + \beta_p x_i^p + \varepsilon_i \qquad i = 1, ..., n \tag{1.1}$$

Avec les hypothèses suivantes :
- Les termes x^j sont supposés déterministes.
- Les ε_i sont des termes d'erreur, non observés, indépendants et identiquement distribués ; $E(\varepsilon_i) = 0$ et $Var(\varepsilon_i) = \sigma^2$.
- Les paramètres inconnus $\beta_1, \beta_2, ..., \beta_p$ sont supposés constants.
- Les erreurs sont linéairement indépendantes à des variables exogènes (ie $\text{cov}(X_i, \varepsilon_j) = 0 \ \forall i \neq j$).

  L'écriture matricielle de (1.1) nous donne la définition suivante :

Définition 1.1.1. ***(Modèle de régression multiple)***
Un modèle de régression linéaire multiple est défini par une équation de la forme

$$Y_{n\times 1} = X_{n\times(p+1)}\beta_{(p+1)\times 1} + \varepsilon_{n\times 1} \tag{1.2}$$

où :
- *Y est le vecteur des n valeurs observées de la variable à expliquer.*
- *X est la matrice des données à n lignes et $p+1$ colonnes.*
- *β est le vecteur de dimension p des paramètres inconnus du modèle.*
- *ε est le vecteur de dimension n des erreurs.* ■

Nous faisons l'hypothèse supplémentaire que la matrice $X'X$ est inversible, c'est-à-dire que la matrice X est de rang $(p+1)$ et donc qu'il n'existe pas de colinéarité entre ses colonnes. En pratique, si cette hypothèse n'est pas vérifiée, il suffit de supprimer des colonnes de X et donc des variables du modèle.

1.1.2 Estimateur des moindres carrés

Définition 1.1.2. ***(Estimateur des MC)***
On appelle estimateur des moindres carrés (noté MC) $\hat{\beta}$ de β la valeur suivante :

$$\hat{\beta} = arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_i^j \right)^2 = arg\min_{\beta} (Y - X\beta)'(Y - X\beta)$$

Théorème 1.1.1. ***(Expression de l'estimateur des MC)***
L'estimateur des MC $\hat{\beta}$ de β vaut :

$$\hat{\beta} = (X^{'}X)^{-1}X^{'}Y.$$

Preuve

L'expression à minimiser sur $\beta \in \mathbb{R}^{p+1}$ s'écrit :

$$\begin{aligned}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i^1 - \beta_2 x_i^2 - ... - \beta_p x_i^p\right)^2 &= \| Y - X\beta \|^2 \\ &= (Y - X\beta)^{'}(Y - X\beta) \\ &= Y^{'}Y - 2\beta^{'}X^{'}Y + \beta^{'}X^{'}X\beta.\end{aligned}$$

Par dérivation matricielle de la dernière équation on obtient :

$$X^{'}Y - X^{'}X\beta = 0.$$

Alors :

$$\hat{\beta} = (X^{'}X)^{-1}X^{'}Y.$$

Pour s'assurer que ce point $\hat{\beta}$ est bien un minimum strict, il faut que la dérivée seconde soit une matrice définie positive. Or la dérivée seconde sécrit $2X^{'}X$ et X est de plein rang donc $X^{'}X$ est inversible et n'a pas de valeur propre nulle. La matrice $X^{'}X$ est donc définie.

De plus $\forall z \in \mathbb{R}^p$, nous avons :

$$z^{'}2X^{'}Xz = 2\langle Xz, Xz\rangle = 2||Xz|| \geq 0.$$

$(X^{'}X)$ est donc bien définie positive et $\hat{\beta}$ est bien un minimum strict. ■

1.1.3 Quelques propriétés statistiques

Proposition 1.1.1. *L'estimateur $\hat{\beta}$ des MC est un estimateur sans biais de β et sa variance vaut $V(\hat{\beta}) = \sigma^2(X^{'}X)^{-1}$.*

Preuve

$$\mathbb{E}(\hat{\beta}) = \mathbb{E}((X^{'}X)^{-1}X^{'}Y) = (X^{'}X)^{-1}X^{'}\mathbb{E}(Y) = (X^{'}X)^{-1}X^{'}X\beta = \beta.$$

L'estimateur des MC est donc sans biais.

Calculons sa variance :

$$V(\hat{\beta}) = V((X^{'}X)^{-1}X^{'}Y) = (X^{'}X)^{-1}X^{'}V(Y)X(X^{'}X)^{-1} = \sigma^2(X^{'}X)^{-1}.\blacksquare$$

Le théorème de Gauss-Markov nous indique que parmi tous les estimateurs linéaires sans biais de β, l'estimateur obtenu par MC admet la plus petite variance :

Théorème 1.1.2. ***(Gauss-Markov)****[31]*
L'estimateur $\hat{\beta}$ des MC est optimal parmi les estimateurs linéaires sans biais de β.

1.1.4 Valeurs ajustées et vecteur des résidus

Les valeurs ajustées (ou estimées, prédites) de Y ont pour expression :

$$\hat{Y} = X\hat{\beta} = X(X^{'}X)^{-1}X^{'}Y = \mathbf{H}Y,$$

où $\mathbf{H} = X(X^{'}X)^{-1}X^{'}$ est appelée *"hat matrix"* ; elle met un chapeau à Y. Géométriquement, c'est la matrice de projection orthogonale dans $\mathbb{R}^n$ sur le sous-espace $\text{Vect}(X)$ engendré par les vecteurs colonnes de X.

On note :

$$e = Y - \hat{Y} = Y - X\hat{\beta} = (\mathbf{I} - \mathbf{H})Y.$$

Le vecteur des résidus ; c'est la projection de Y sur le sous-espace orthogonal de $\text{Vect}(X)$ dans $\mathbb{R}^n$.

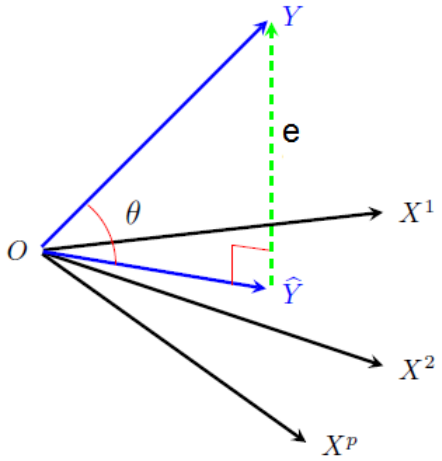

Figure 1.1- Géométriquement, la régression est la projection $\hat{Y}$ de Y sur l'espace vectoriel $Vect(\mathbb{1}, X^1, ..., X^p)$.

1.2 Problème de la multicolinéarité

1.2.1 Définition et conséquences

Définition 1.2.1. *La multicolinéarité est le fait qu'une variable explicative est une combinaison linéaire des autres variables explicatives. Par exemple si une variable X_3 en faisant la somme pondérée de deux autres variables X_1 et X_2, par exemple $X_3 = 2X_1 + 3X_2$, alors X_1, X_2 et X_3 seront multicolinéaires et on parle de multicolinéarité parfaite.*

En pratique on rencontre rarement ce genre de situation. Par contre, on rencontre assez souvent la situation où X_3 est trés proche d'une combinaison linaire de X_1 et X_2. Cela veut dire que la régression de X_3 sur X_1 et X_2 est très significative. Dans ce cas la variable X_3 partage une partie de sa variabilité avec X_1 et X_2.

Conséquences

- Si la multicolinéarité est parfaite alors la matrice $(X'X)^{-1}$ n'est pas inversible et donc l'estimateur MC n'est pas calculable.
- Lorsque l'une des variables explicatives est proche d'une combinaison linaire des

autres variables alors $X^{'}X$ serait mal conditionnée ($\det(X^{'}X)$ proche de 0) $(X^{'}X)^{-1}$ aura des éléments très grands.

– En cas de multicolinéarité, certaines des variances estimées des coefficients vont être trés grands.
– L'écart-type $\hat{\sigma}_{\beta_i}$ d'un coefficient β_i est un indicateur de la stablité de l'estimation de ce dernier. Si $\hat{\sigma}_{\beta_i}$ est du même ordre de grandeur que β_i, ce dernier est mal déterminé.

1.2.2 Comment détecter la multicolinéarité

– **Critère de Klein** : Il s'agit simplement d'un critère de présomption de la multicolinéarité. Il y a présomption de multicolinéarité si au moins un des $r(X_i, X_j)$ élevé au carré est supérieur au R^2 [1] (le coefficient de détermination de la régression).
– **Le test de Farrar et Glauber** : Teste les hypothéses $\mathcal{H}_0$: Absence de multicolinéartié et $\mathcal{H}_1$: Présence de multicolinéartié. Soit R la matrice des corrélations des variables explicative. Ce test est basé sur la statistique suivante :

$$\mathcal{X}^2 = -(n - 1 - (2p + 7)6ln(|R|).$$

Il y présomption de multicolinéarité si $\mathcal{X}^2 > \mathcal{X}^2_{1-\alpha:p(p+1)/2}$.

1.2.3 Comment corriger le probléme de multicolinéarité

– Utiliser une des procédure de sélection de modéle pour choisir un modéle contenant moins de variables.
– Effectuer une ACP sur les variables explicatives et utiliser les premières composantes principales comme variables explicative.
– Régression Ridge.
– Régression PLS.

1. Le coefficient de détermination R^2 est défini par

$$R^2 = \frac{\|\hat{Y} - \bar{Y}\mathbb{1}\|^2}{\|Y - \bar{Y}\mathbb{1}\|^2}$$

.

Chapitre 2

Régression sur composantes principales (PCR)

Nous avons vu dans le premier chapitre, que la régression linéaire reposait sur l'hypothèse :

$\mathcal{H}_1 : rang(X) = p.$

Nous allons maintenant traiter le cas où $\mathcal{H}_1$ n'est plus vérifiée. Cela revient à dire que $X^{'}X$ a un déterminant nul et donc qu'elle n'est plus inversible.

2.1 Analyse en composantes principales (ACP)

Dans la plupart des situations, on dispose de plusieurs observations sur chaque individu constituant la population d'étude. On a donc pris en comptes p variables par individu.

L'étude séparée de chacune de ces variables donne quelques informations mais insuffisantes car elle laisse de cote les liaisons entre elles, ce qu'est pourtant souvent ce qui l'on veut étudier.

l'ACP est alors une bonne méthode pour étudier les données multidimensionnelles lorsque les variables observées sont du type numérique.

2.1.1 Présentation d'ensembles

Les observations de p variables sur n individus sont rassemblées dans un tableau rectangulaire X à n lignes et p colonnes :

$$X = \begin{pmatrix} x_1^1 & x_1^2 & . & x_1^j & . & x_1^p \\ x_2^1 & x_2^2 & . & x_2^j & . & x_2^p \\ . & . & . & . & . & . \\ x_i^1 & x_i^2 & . & x_i^j & . & x_i^p \\ . & . & . & . & . & . \\ x_n^1 & x_n^2 & . & x_n^j & . & x_n^p \end{pmatrix} \in \mathcal{M}_{(n,p)}(\mathbb{R})$$

x_i^j : est la mesure de la variable x^j sur l'individu x_i, $i = 1, ..., n$ et $j = 1, ..., p$.

Dans la suite, on considère $X_{(n,p)}$ comme un tableau des données centrés et réduits.

$$X = \begin{pmatrix} \frac{x_1^1-\bar{x}^1}{\sigma_{x^1}} & \frac{x_1^2-\bar{x}^2}{\sigma_{x^2}} & . & \frac{x_1^j-\bar{x}^j}{\sigma_{x^j}} & . & \frac{x_1^p-\bar{x}^p}{\sigma_{x^p}} \\ \frac{x_2^1-\bar{x}^1}{\sigma_{x^1}} & \frac{x_2^2-\bar{x}^2}{\sigma_{x^2}} & . & \frac{x_2^j-\bar{x}^j}{\sigma_{x^j}} & . & \frac{x_2^p-\bar{x}^p}{\sigma_{x^p}} \\ . & . & . & . & . & . \\ \frac{x_i^1-\bar{x}^1}{\sigma_{x^1}} & \frac{x_i^2-\bar{x}^2}{\sigma_{x^2}} & . & \frac{x_i^j-\bar{x}^j}{\sigma_{x^j}} & . & \frac{x_i^p-\bar{x}^p}{\sigma_{x^p}} \\ . & . & . & . & . & . \\ \frac{x_n^1-\bar{x}^1}{\sigma_{x^1}} & \frac{x_n^2-\bar{x}^2}{\sigma_{x^2}} & . & \frac{x_n^j-\bar{x}^j}{\sigma_{x^j}} & . & \frac{x_n^p-\bar{x}^p}{\sigma_{x^p}} \end{pmatrix} \in \mathcal{M}_{(n,p)}(\mathbb{R})$$

2.1.2 Nuages des points initiaux

Nuage d'individus

Nous représentons graphiquement les individus par un nuage de points dans un sous-espace E de $\mathbb{R}^p$ et l'information intéressante pour l'individu est la distance entre les points. On définit sur E la métrique $M = \mathbf{I_p}$.

Nuage de variables

Nous représentons graphiquement les variables par un nuage de points dans un sous-espace F de $\mathbb{R}^n$ et on définit la métrique de poids $D = \frac{1}{n}\mathbf{I_n}$.

2.1.3 Principe de l'ACP :

On cherche à définir k nouvelles variables combinaisons linéaires des p variables initiales qui feront perdre le moins d'information possible.

– Ces variables seront appelées «**composantes principales** »
– Les axes qu'elles déterminent : « **axes principaux** »
– Les formes linéaires associées : « **facteurs principaux** »

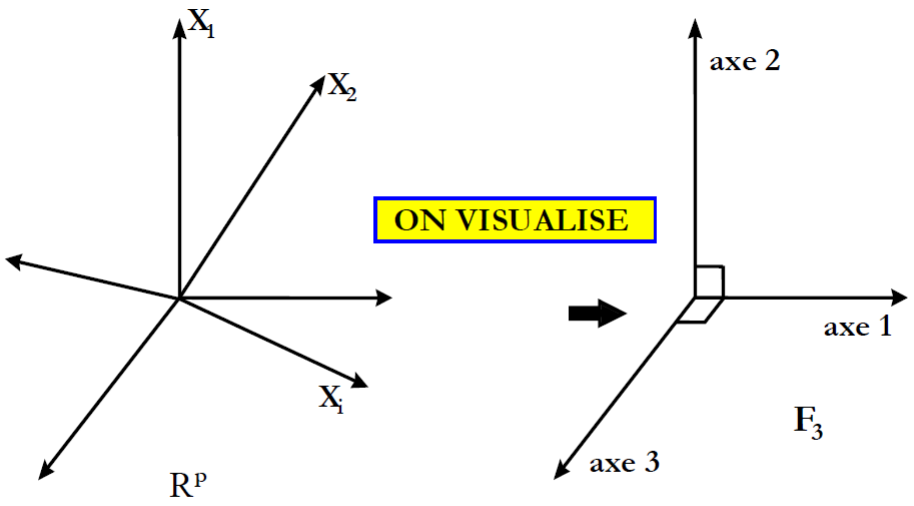

Figure (2.1) Les k nouveaux axes.

Pour perdre le moins d'information possible il faut que :

1. F_k devra être « ajusté » le mieux possible au nuage des individus : la somme des carrés des distances des individus à F_k doit être minimale.
2. F_k est le sous-espace tel que le nuage projeté ait une inertie [1] (dispersion) maximale.

1. L'inertie est la somme pondérée des carrés des distances des individus au centre de gravité g.

$$I_g = \sum_{i=1}^{n} \frac{1}{n} d^2\left(X_i, \mathrm{g}\right),$$

ou de façon plus générale :

$$I_g = \sum_{i=1}^{n} p_i d^2\left(X_i, \mathrm{g}\right) \qquad \text{avec} \quad \sum_{i=1}^{n} p_i = 1$$

2.1.4 Équivalence des deux critères concernant la perte d'information

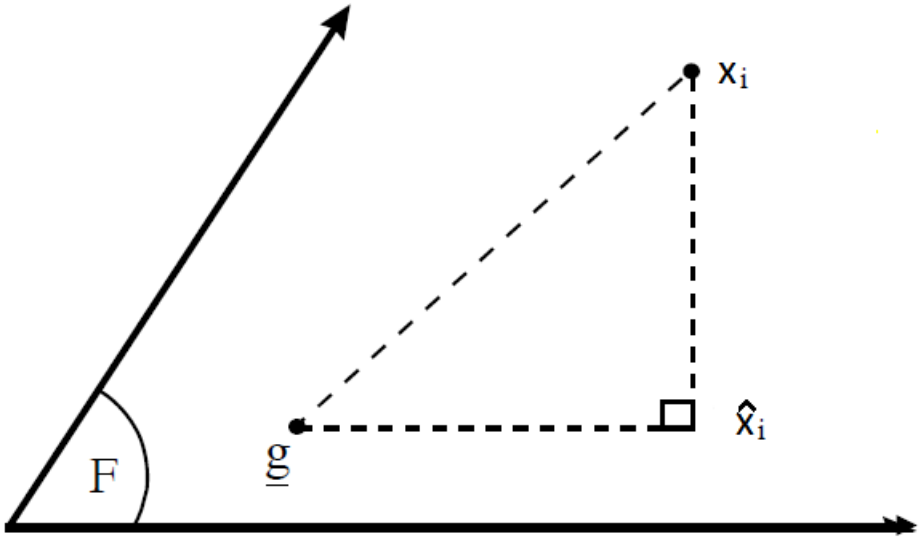

Figure (2.2) Projection orthogonale de X_i sur F.

Soit F un sous-ensemble de $\mathbb{R}^p$.
$\hat{X}_i$ la projection orthogonale de X_i sur F.

$$\|X_i - \mathrm{g}\|^2 = \|X_i - \hat{X}_i\|^2 + \|\hat{X}_i - \mathrm{g}\|^2 \qquad \forall i = 1, ..., n$$

On va chercher F tel que :

$$\sum_{i=1}^{n} p_i \|X_i - \hat{X}_i\|^2 \text{ soit minimal.}$$

Ce qui revient d'après le théorème de Pythagore à maximiser :

$$\sum_{i=1}^{n} p_i \|\hat{X}_i - \mathrm{g}\|^2.$$

2.1.5 Éléments principaux de l'ACP :

Les axes principaux

Les axes principaux sont les droites engendrées par les k premiers vecteurs propres associés au k premières valeurs propres de la matrice $\Gamma = VM$ tel que $V = X^{'}DX$.

Les facteurs principaux

Les facteurs principaux sont les vecteurs propres associés aux valeurs propres de la matrice MV.

Les composantes parincipales

Les composantes principales sont les vecteurs de base de l'espace F_k ($F_k \subset F$ l'espace le plus proche au nuage des variables projetés), notées par $\{C_1, ..., C_k\}$:

$$C_i = XMu_i \quad \forall i \in \{1, ..., n\}$$

Les composantes principales sont les vecteurs propres associés aux valeurs propres de la matrice WD avec $W = XMX^{'}$.

Propriétés des composantes principales

1. La variance d'une composante principale est égale à l'inertie portée par l'axe principal qui lui est associé qui est égale à λ.
2. Les composantes principales sont non corrélées deux à deux.
   En effet, les axes associés sont orthogonaux.

2.2 Régression sur composantes principales

Cette méthode consiste à introduire un changement de variable afin de reparamétriser le problème de régression et introduire les composantes principales. La matrice $(X^{'}X)$ est une matrice symétrique, nous pouvons donc écrire :

$$X^{'}X = P\Lambda P^{'} \tag{2.1}$$

où P est la matrice des vecteurs propres normalisés de ($X^{'}$X), c'est-à-dire que P est une matrice orthogonale ($P^{'}P = \mathbf{I}$) et $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_p)$ est la matrice diagonale des valeurs propres classées par ordre décroissant $\lambda_1 \geq \lambda_2 \geq .... \geq \lambda_p$.

Remarque 2.2.1. *Si l'on effectue l'analyse en composantes principales (ACP) du tableau X (ou du triplet $(X, I_p, In/n)$), la matrice P est la matrice des axes principaux normés à l'unité, mais les valeurs propres de l'ACP sont les $\{\lambda_j\}$ avec j variant de 1 à p divisés par n.*

2.2.1 Hypothèse $\mathcal{H}_1$ satisfaite : $|X^{'}X| \neq 0$

La matrice X est de plein rang. Analysons l'impact de la transformation précédente sur le modèle de régression qui s'écrit alors :

$$\begin{aligned} Y &= X\beta + \varepsilon \\ &= XP^{'}P\beta + \varepsilon. \end{aligned}$$

$$Y = X^{*}\beta^{*} + \varepsilon. \tag{2.2}$$

où $X^{*} = XP$ correspond aux composantes principales, $X_i^{*} = XP_i$ et $X_i^{*'}X_i^{*} = \lambda_i$. Lors de l'ACP du tableau (ou du triplet $(X, I_p, In/n)$), les composantes principales normées à la valeur propre obtenues sont égales aux vecteurs X_i^{*} que l'on obtient ici, d'où le nom de la méthode. Cette dernière equation (2.2) définit un modèle de régression que nous appellerons modèle « étoile » qui est tout simplement la régression sur les composantes principales X^{*}. Remarquons de plus que par construction :

$$X^{*'}X^{*} = P^{'}X^{'}XP = P^{'}P\Lambda P^{'}P = \Lambda \tag{2.3}$$

Les nouvelles variables de X^* sont orthogonales et de norme λ_j par construction c'est une propriété classique des composantes principales d'une ACP.

La solution classique des MC vaut $\hat{\beta} = (X'X)^{-1}X'Y$ et la variance de cet estimateur vaut :

$$V(\hat{\beta}) = \sigma^2(X'X)^{-1}.$$

Si maintenant nous calculons l'estimateur des MC dans le modèle « étoile », c'est-à-dire si nous effectuons une régression sur les composantes principales, nous obtenons :

$$X^*\hat{\beta}^* = X^*(X^{*'}X^*)^{-1}X^{*'}Y,$$

qui peut s'écrire sous la forme simplifiée suivante :

$$X^*\hat{\beta}^* = XP\Lambda^{-1}P'X'Y.$$

L'estimateur du modèle étoile est donc :

$$\hat{\beta}^* = \Lambda^{-1}P'X'Y.$$

Cet estimateur minimise les moindres carrés puisque les moindres carrés du modèle étoile et du modèle initial sont identiques par construction :

$$\|Y - X\beta\|^2 = \|Y - XPP'\beta\|^2 = \|Y - X^*\beta^*\|^2.$$

En utilisant (2.3), la variance de cet estimateur vaut :

$$V(\hat{\beta}^*) = \sigma^2(X^{*'}X^*)^{-1} = \sigma^2\Lambda^{-1}.$$

Les estimateurs des coefficients de chacune de ces nouvelles variables explicatives sont non corrélés. La variance pour l'estimation du coefficient de la i$^{\text{ième}}$ variable X_i^* est $\sigma^2\lambda_i^{-1}$. Pour $i < j$ nous avons $V(\hat{\beta}_i^*) < V(\hat{\beta}_j^*)$, cela veut dire que l'estimation est plus précise sur les premières composantes principales de X.

Comme les composantes principales sont orthogonales entre elles, l'estimation des β_i^* peut se faire par régression linéaire simple sans constante sur la i$^{\text{ième}}$ composante principale X_i^*.

2.2.2 Colinéarité parfaite : $|X^{'}X| = 0$

Reprenons l'équation (2.1)

$$X^{'}X = P\Lambda P^{'}.$$

Le rang de X vaut maintenant k avec $k < p$, nous avons donc les $(p-k)$ dernières valeurs propres de $(X^{'}X)$ qui valent zéro, $\lambda_{k+1} = ... = \lambda_p = 0$. Cela veut dire que pour tout $i > k$, nous avons :

$$X_i^{*'}X_i^* = \lambda_i = 0. \tag{2.4}$$

Décomposons la matrice Λ en matrices bloc :

$$\Lambda = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \Lambda_1 = diag\left(\lambda_1, ..., \lambda_k\right),$$

et décomposons la matrice orthogonale P de taille $p \times p$ qui regroupe les vecteurs propres normés de $X^{'}X$ en deux matrices P_1 et P_2 de taille respectivement $p \times k$ et $p \times (p-k)$. Soit $P = [P_1, P_2]$, nous avons alors :

$$X^* = [X_1^*, X_2^*] = [XP_1, XP_2]\ .$$

Cherchons maintenant la valeur de XP_2. Comme le rang de X vaut k, nous savons que la dimension de $F(X)$ vaut k et de même pour la dimension de $F(X^{'}X)$. Ce sous-espace vectoriel possède une base à k vecteurs que l'on peut choisir orthonormés. Nous savons, par construction, que P_1 regroupe k vecteurs de base orthonormés de $F(X^{'}X)$ tandis que P_2 regroupe $(p-k)$ vecteurs orthonormés (et orthogonaux aux k de P_1) qui complètent la base de $F(X^{'}X)$ d'obtenir une base de $\mathbb{R}^p$. Nous avons donc que quel que soit $u \in F(X^{'}X)$ alors :

$$u^{'}P_2 = 0.$$

Prenons $u \neq 0$ et comme $u \in F(X^{'}X)$, il existe $\gamma \in \mathbb{R}^p$ tel que $u = X^{'}X\gamma \neq 0$. Nous avons donc :

$$\gamma^{'}X^{'}XP_2 = 0,$$

pour tout $\gamma \in \mathbb{R}^p$ et donc $X^{'}XP_2 = 0$, c'est-à-dire que $XP_2 = 0$. Nous avons alors :

$$X^* = [X_1^*, X_2^*] = [XP_1, XP_2] = [XP_1, 0].$$

Au niveau des coefficients du modèle étoile, nous avons la partition suivante :

$$\beta^* = \begin{pmatrix} \beta_1^* \\ \beta_2^* \end{pmatrix} = \begin{pmatrix} P_1'\beta \\ P_2'\beta \end{pmatrix}$$

Grâce à la reparamétrisation précédente et avec $X_2^* = XP_2 = 0$, nous avons :

$$\begin{aligned} Y &= X^*\beta^* + \varepsilon \\ &= X_1^*\beta_1^* + X_2^*\beta_2^* + \varepsilon \\ &= X_1^*\beta_1^* + \varepsilon. \end{aligned}$$

Cette paramétrisation nous assure donc que les moindres carrés dans le modèle initial et dans le modèle étoile sont égaux et nous allons donc utiliser le modèle étoile. Par les MC, nous obtenons $\beta_1^* = \left(X_1^{*'}X_1^*\right)^{-1} X_1^{*'}Y$ et nous posons $\beta_2^* = 0$ ce qui ne change rien car $X_2^* = 0$. Nous obtenons l'estimateur de la régression sur les k premières composantes principales (*principal component regression*) :

$$\hat{\beta}_1^* = \Lambda_1^{-1} P_1' X^{'} Y,$$

de variance :

$$V(\hat{\beta}_1^*) = \sigma^2 (X_1^{*'}X_1^*)^{-1} = \sigma^2 \Lambda_1^{-1}. \tag{2.5}$$

La stabilité des estimateurs peut être envisagée par leur variance, plus celle-ci est grande, plus l'estimateur sera instable. Cette variance dépend ici du bruit qui fait partie du problème et de λ_j. Une très faible valeur propre induit une grande variance et donc un estimateur instable et des conclusions peu fiables. Nous avons donc que $\hat{\beta}_1^*$ minimise le critère des MC pour le modèle étoile.

Comme les MC du modèle étoile et ceux du modèle initial sont égaux, à partir de $\hat{\beta}_1^*$, le vecteur des coefficients associés aux composantes principales, nous pouvons obtenir simplement :

$$\hat{\beta}_{PCR} = P_1 \hat{\beta}_1^*.$$

Ce vecteur de coefficient minimise les MC du modèle initial. Le résultat est donc identique au paragraphe précédent à ceci près que l'on s'arrête aux k premières composantes principales associées aux valeurs propres non nulles de $(X'X)$.

Ceci suggère le fait que l'on peut trouver une valeur, pour l'estimateur de la régression $\hat{\beta}$ qui est égale à $\hat{\beta}_1^*$. Mais nous pourrions trouver une infinité d'autres $\hat{\beta}$ qui seraient aussi solution de la minimisation des MC. Ils seraient tels que $\hat{\beta}_2 \neq 0$. Ceci donnerait un estimateur :

$$\hat{\beta} = P_1 \hat{\beta}_1 + P_2 \hat{\beta}_2.$$

En plaçant cette valeur dans les moindres carrés cela donne exactement les mêmes moindres carrés que ceux obtenus par $\hat{\beta}_{PCR}$. Nous retrouvons là le fait que $\hat{\beta}$ n'est plus unique car $\mathcal{H}_1$ n'est plus vérifiée. Par contre, nous avons que $\hat{\beta}_{PCR}$ est unique. Puisque les résultats sont conservés quand l'on s'arrête à k, ce paragraphe suggère aussi que nous pouvons choisir une valeur de k de sorte que les valeurs propres associées $\{\lambda_j\}_{j=1}^k$ soient suffisamment différentes de 0, éliminant ainsi les problèmes de quasi non-inversibilité et de variance très grande. C'est cette méthode que nous allons exposer dans le prochain paragraphe. Evidemment, si l'on élimine les composantes principales associées à des valeurs propres non strictement nulles voire suffisamment grandes, la solution des MC dans le modèle initial et celle dans le modèle étoile seront différentes. Cependant, dans l'approche régression sur composantes principales, nous ne garderons que les estimateurs stables (ie, de faible variance). Cette différence de moindres carrés est le prix à payer afin d'obtenir une solution unique et stable.

2.2.3 Pratique de la régression sur composantes principales

Nous utilisons la paramétrisation du problème (2.2) et nous avons donc :

$$Y = X^* \beta^* + \varepsilon,$$

où $X^* = XP$ représente la matrice des p composantes principales, P représente la matrice des p vecteurs propres normés à l'unité de la matrice $X'X$ (ou axes principaux) associés aux valeurs propres $(\lambda_1, \lambda_2, ..., \lambda_p)$ classées par ordre décroissant. Nous sommes donc en présence de p nouvelles variables (les composantes principales) qui sont orthogonales entre elles. Si l'on conserve toutes les composantes principales, le résultat est identique à la régression classique, un changement de variable mis à part.

Le but de la régression sur composantes principales consiste à ne conserver qu'une partie des composantes principales, à l'image de ce qui est fait en analyse en composantes principales (ACP). Les k composantes principales conservées seront la part conservée de l'information contenue dans les variables explicatives, alors que les $(p-k)$ éliminées seront la part d'information contenue dans les variables explicatives qui sera éliminée, car considérée comme négligeable. Ici l'information est mesurée en terme d'inertie ou de dispersion et est égale à la valeur propre : plus la valeur propre λ_j est élevée, plus la part d'information apportée par la composante j est importante propos illustrés par l'équation (2.4). Il semble donc assez naturel de ne conserver que les composantes dont la part d'information associée est grande, à savoir conserver les composantes associées aux k premières valeurs propres. Les estimateurs des coefficients des k premières composantes principales retenues seront les moins variables (2.5). Les étapes d'une régression sur composantes principales sont données ci-dessous :

Centrage-réduction

À la différence de la régression classique où les variables sont en général conservées telles que mesurées, il est d'usage de centrer et réduire toutes les variables au préalable, tant les p variables explicatives que la variable à expliquer Y. Une variable centrée-réduite $\bar{X}_j$ issue de la variable X_j s'écrit donc :

$$\bar{X}_j = (X_j - \bar{X}_j \mathbb{1}_n)/\hat{\sigma}_j,$$

où $\bar{X}_j$ est la moyenne empirique de X_j et $\hat{\sigma}_j^2$ un estimateur de la variance.

Cette pratique a pour but d'accorder la même importance pour le choix des composantes. En effet, si deux variables explicatives sont mesurées à des échelles telles

que la première varie de 10^{-3} (par exemple un poids en tonne) autour de sa moyenne et la seconde varie de 10^5 autour de sa moyenne (par exemple des âges mesurés en heures), alors la composante va privilégier la direction ayant le maximum de dispersion, c'est-à-dire l'âge, et ce juste pour un problème d'unité. Cette étape est donc en général nécessaire.

Après centrage-réduction, nous avons que le produit scalaire entre deux variables centrées-réduites est la corrélation linéaire r :

$$\langle \bar{X}_j, \bar{X}_l \rangle = r(X_j, X_l).$$

De plus, les composantes principales sont de norme λ_j et orthogonales entre elles. Ces composantes sont des vecteurs de $\mathbb{R}^n$ qui sont des variables « synthétiques » constituées par une combinaison linéaire des variables initiales car $X^* = \bar{X}P$. Nous avons donc pour la j$^{\text{ième}}$ composante principale la relation :

$$X_j^* = \bar{X}P_j,$$

donc sa moyenne empirique $\bar{X}_j^*$ vaut 0. Le produit scalaire entre deux composantes principales est donc la covariance empirique et l'équation (2.3) se traduit simplement comme « les composantes principales sont non corellées entre elles » et de variances décroissantes égales à λ_j.

Choix du nombre de composantes k du modèle

Le problème délicat de la régression sur composantes principales est la détermination du nombre de composantes k à conserver.

1- Méthode graphique

Pour déterminer k, il est possible, à l'image de ce qui est fait en ACP, de tracer le diagramme en tuyaux d'orgue des valeurs propres et de choisir le numéro k de la valeur propre après laquelle les valeurs propres sont nettement plus petites. En général, cette procédure est adaptée à l'interprétation (c'est-à-dire à l'ACP), mais sélectionne trop peu de composantes pour un modèle utilisé à des fins de prévision.

2- Apprentissage-validation

La procédure de validation consiste à séparer de manière aléatoire les données en deux parties distinctes (X_a, Y_a) et (X_v, Y_v). Le cas échéant le jeu d'apprentissage est centré-réduit. Les valeurs des moyennes et des variances serviront à calculer les prévisions sur les données de validation. Une régression sur composantes principales est conduite avec le jeu d'apprentissage (X_a, Y_a) pour tous les nombres de composantes principales possibles. Ensuite, en utilisant tous ces modèles et les variables explicatives X_v, les valeurs de la variable à expliquer sont prédites $\hat{Y}_v^{PCR}(k)$ pour tous les k. Si le modèle est estimé sur des données centrées-réduites, la prévision des données initiales s'obtient à partir du modèle centré-réduit par :

$$\hat{Y}_v^{PCR}(k) = \hat{\sigma}_{aY}^2 \sum_{j=1}^{p} \frac{X_{vj} - \bar{X}_{aj}\mathbb{1}_{nv}}{\hat{\sigma}_{aj}} \hat{\beta}_j^*(k) + \hat{Y}_a \mathbb{1}_{nv}.$$

La qualité du modèle est ensuite obtenue en mesurant la distance entre les observations prévues et les vraies observations par un critère. Le plus connu est le PRESS :

$$PRESS(k) = ||\hat{Y}_v^{PCR}(k) - Y_v||^2.$$

D'autres critères peuvent être utilisés comme :

$$MAE(k) = ||\hat{Y}_v^{PCR}(k) - Y_v||_1,$$

où $||x||_1 = \sum_i |x_i|$ est la norme de type l^1.
Le nombre de composantes principales optimal k choisi est celui qui conduit à la minimisation du critère choisi.

3-Validation croisée

Il est aussi possible de choisir k par validation croisée. Pour toutes les valeurs de k possibles (k variant de 1 à K fixé, avec $K \leq rang(X)$), on supprime une observation (ou un groupe de b observations) puis on estime le modèle sans cette (ou ces) observation(s). On peut alors prévoir cette (ou ces) observation(s) grâce à ce modèle estimé. Dans le cas d'une seule observation enlevée, la i$^{\text{ième}}$, pour un nombre

de composantes k, la prévision est notée $\hat{y}_{(i)}(k)$. On peut enfin à l'aide d'un critère, par exemple le PRESS, connaître la capacité de prévision d'un modèle à k composantes par :

$$PRESS(k) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_{(i)}^{PCR}(k)\right)^2.$$

Le nombre optimal k de composantes est celui qui réalise le minimum du PRESS.

Calculs et estimations

Une fois k choisi, les calculs sont identiques à ceux de la section 2.2.1. Le logiciel utilisé nous permet de calculer les k premiers axes principaux du tableau centré-réduit $\bar{X}$. Ils sont regroupés dans la matrice P_1 orthogonale. À partir de cette matrice, sont calculées les composantes principales $X_1^* = XP_1$. Nous pouvons en déduire l'estimateur des coefficients associés aux composantes principales :

$$\hat{\beta}_1^* = \left(X_1^{*'}X_1^*\right)^{-1} X_1^{*'}Y.$$

Bien entendu, la valeur de k est choisie inférieure ou égale au rang de $X^{*'}X^*$ et donc l'inverse $\left(X^{*'}X^*\right)^{-1}$ existe. Comme il est souvent difficile d'accorder une signification précise aux composantes principales, nous pouvons réexprimer les coefficients en fonction des variables initiales (centrées-réduites) :

$$\hat{\beta}_{PCR} = P_1\hat{\beta}_1.$$

Les variances des estimateurs sont égales à :

$$V(\hat{\beta}_1^*) = \sigma^2(X_1^{*'}X_1^*)^{-1} = \sigma^2\Lambda_1^{-1}, \tag{2.6}$$

$$V(\hat{\beta}_{PCR}^*) = \sigma^2 P_1\Lambda_1^{-1}P_1'.$$

Enfin, le modèle permet de faire la prévision, il suffit d'utiliser le modèle étoile. Cela donne :

$$\hat{Y}_{PCR} = X\beta_{PCR}.$$

Comme le modèle étoile est un modèle de régression, tous les résultats s'appliquent à ce modèle.

Remarque 2.2.2. *En général $k < rang(X)$ et donc les moindres carrés obtenus avec la régression linéaire et ceux obtenus avec la régression sur composantes principales sont différents et les coefficients nont aucune raison d'être identiques. Il s'agit de deux modélisations différentes.*

Conclusion

L'avantage de la régression en composantes principales est de conserver une partie de l'information et d'utiliser de nouvelles variables qui sont orthogonales. Il en résulte une simplicité de calcul et une stabilité des estimations si k est convenablement choisi. Les composantes étant orthogonales, les tests de nullité de coefficients β_j^* associés aux composantes principales X_j^* (indépendantes les unes des autres) s'effectuent facilement.

Un inconvénient de la régression sur composantes principales réside dans le choix de k et un autre dans l'interprétation des variables. En effet, les nouvelles variables ne sont pas toujours interprétables puisqu'elles sont des combinaisons linéaires des variables explicatives originales. Cela est toutefois un inconvénient mineur car nous pouvons revenir aux variables initiales via $\hat{\beta}_{PCR}$. Le retour aux variables initiales fait tout de même perdre la propriété d'orthogonalité des variables. Le principal inconvénient réside dans l'élimination des $(p-k)$ composantes principales de faibles variances (ou inerties), or ce sont peut-être ces composantes de faibles variances qui sont les plus explicatives.

Cette méthode n'est plus très utilisée actuellement, il est peut-être préférable d'utiliser une régression partial least square (PLS), qui conserve les mêmes avantages mais qui choisit des composantes en tenant compte de leur covariance avec la variable Y à expliquer.

Chapitre 3

Régression aux moindres carrés partiels (PLS)

3.1 Modèle

Comme dans la régression linéaire multiple, le but principal de la régression PLS est de construire un modèle linéaire :

$$Y = X\beta + \varepsilon,$$

où $\beta_{p\times c}$ c coeffcients de régression, $\varepsilon_{n\times c}$ terme de bruit pour le modéle.

Usuellement les variables dans X et Y sont centrées et réduites. La régression sur composantes principales et la régression PLS produisent toutes les deux des facteurs de scores comme des combinaisons linéaires des variables prédictives originelles, de telle manière qu'il n'y ai pas de corrélation entre les facteurs scores utilisés par le modèle de régression prédictive.

Par exemple, supposons que nous ayons un ensemble de données pour des variables réponses Y et un grand nombre de variables prédictives X, dont certaines sont fortement corrélées. Une régression utilisant l'extraction des facteurs pour ce type de données calcule la matrice de facteurs score $T = XW$ pour une matrice de poids appropriée W et alors on considère le modéle de régression linéaire $Y = TQ + \varepsilon$ où Q est une matrice des cofficients de régression pour T et ε un terme d'erreur. Une

fois les Q calculées, le modèle de régression ci-dessus est équivalent à $Y = X\beta + \varepsilon$ où $\beta = WQ$ qui peut être utilisé comme un modèle de regression prédictive.

Remarque 3.1.1. *La régression sur composantes principales et la régression PLS diffèrent dans la méthode utilisée pour extraire les facteurs de scores. En résumé, la régression sur composantes principales produit une matrice de poids W reflétant la structure de covariances entre les variables prédictives alors que la régression PLS produit une matrice de poids W reflétant les structures de covariance entre les prédicteurs et les réponses.*

Pour établir le modèle, la régression PLS produit une matrice de poids $W_{p\times c}$ pour X telle que $T = XW$, c'est-à-dire les colonnes de W sont des vecteurs de poids pour les colonnes de X produisant la matrice de facteurs de score $T_{n\times c}$ correspondante. Ces poids sont calculés de telle façon qu'ils maximisent la covariance entre la réponse et les facteurs de score correspondants.

La procédure OLS (Ordinary Least Squares) pour le régression de Y sur T est alors utilisée pour produire Q, tel que $Y = TQ + \varepsilon$.

Une fois Q calculé, nous avons $Y = X\beta + \varepsilon$ où $\beta = WQ$ et le modèle de prédiction est complet.

Une matrice supplémentaire nécessaire pour une description complète de la procédure de régression PLS est la matrice P des facteurs qui donne le modèle $X = TP + F$ où F est la partie non expliquée du score de X.

3.2 Méthode

La méthode PLS est une méthode de régression linéaire de c variables réponses sur p variables explicatives toutes mesurées sur les mêmes n individus. Les tableaux des observations, notés respectivement Y et X, de dimensions $n \times c$ et $n \times p$, sont supposés centrés et éventuellement réduits par rapport aux poids $(p_1, ..., p_n)$. On note $D = diag(p_1, ..., p_n)$ la matrice diagonale des poids.

L'intérêt de la méthode comparée à la régression sur composantes principales (PCR), réside dans le fait que les composantes PLS sur les X, notées t, sont calculées

"dans le même temps" que des régressions partielles sont exécutées. Cette simultanéité leur confère un meilleur pouvoir prédictif que celles de la PCR. La question est donc d'examiner comment cette simultanéité est mise en oeuvre.

Notons $E_0 = X$ et $F_0 = Y$ les tableaux centrés et réduits au sens de D qui en général est égal à $\frac{1}{n} I_n$. La méthode procède par étapes succesives permettant le calcul des composantes PLS. On notera K le nombre total d'étapes, c'est-à-dire de composantes indicées par $k = 1, ..., K$.

3.2.1 Description de la k$^{\text{ième}}$ étape

Notons $t = E_{k-1}w$ et $u = F_{k-1}q$, les combinaisons linéaires colonnes des matrices centrées E_{k-1} et F_{k-1}, associées respectivement aux vecteurs des poids w et q. La covariance entre t et u s'écrit comme le D-produit scalaire

$$cov(t, u) = (t, u)_D = w^{'} E_{k-1}^{'} D F_{k-1} q \ .$$

Le carré de la D-norme associée fournit la variance $||t||_D^2 = var(t)$.

L'étape k, se décompose en deux parties. La première fournit les composantes $t_k = E_{k-1}w_k$ et $u_k = F_{k-1}q_k$ par le calcul des poids optimaux w_k et q_k. La deuxième actualise les matrices des prédicteurs et des réponses E_k et F_k comme résidus de la régression sur t_k.

Calcul des poids :

$(w_k, q_k) = argmax cov(t, u) = w^{'} E_{k-1}^{'} D F_{k-1} q$ sous les contraintes $||w||^2 = ||q||^2 = 1$.

Actualisation :

$$E_k = E_{k-1} - P_{t_k} E_{k-1} \quad et \quad F_k = F_{k-1} - P_{t_k} F_{k-1}.$$

où $P_{t_k} = \frac{t_k t_k^k D}{var(t_k)}$ est la matrice $n \times n$ de projection D-orthogonale sur t_k.

Remarquons que le critère à optimiser, la covariance, est un compromis entre le critère de l'Analyse des Corrélations Canoniques, la corrélation, et celui de l'ACP sur chacun des tableaux, la racine carrée de la variance.

$$\begin{aligned} cov(t_k, u_k) &= cov(X_{k-1}w_k, Y_{k-1}q_k). \\ cov^2(X_{k-1}w_k, Y_{k-1}q_k) &= cor^2(X_{k-1}w_k, Y_{k-1}q_k) var(X_{k-1}w_k) var(Y_{k-1}q_k). \end{aligned}$$

3.2.2 Problème d'optimisation

La fonction qu'il faut maximiser est :

$$\begin{aligned} \mathbb{R}^p \times \mathbb{R}^c &\mapsto \mathbb{R} \\ (w,q) \mapsto \phi(w,q) &= cov(t,u) \\ &= t^{'} Du \\ &= u^{'} Dt \end{aligned}$$

Géométriquement :

$$\phi(w,q) = ||t||_D ||u||_D \ ,$$

et statistiquement :

$$\phi(w,q) = \sigma(t)\sigma(u)r(t,u) \quad o\grave{u} \quad \sigma(t) = ||t||_D = \sqrt{t' Dt} \quad et \quad \sigma(u) = ||u||_D = \sqrt{u' Du}$$

Alors la fonction de Lagrange associée à ce problème est :

$$\begin{aligned} \mathbb{R}^p \times \mathbb{R}^c \times \mathbb{R} \times \mathbb{R} &\mapsto \mathbb{R} \\ (w,q,\lambda,\mu) &\mapsto L(w,q,\lambda,\mu) = \phi(w,q) + \frac{\lambda}{2}(1 - ||w||_2^2) + \frac{\mu}{2}(1 - ||q||_2^2) \end{aligned}$$

Le problème de maximisation de ϕ sous contrainte est équivalent au problème de maximisation de L sans contrainte. Pour cela il faut calculer le gradient de L et résoudre le système suivant :

$$\begin{aligned} \nabla_w L &= \nabla_w \phi(w,q) - \frac{\lambda}{2} \nabla_w ||w||_2^2 = 0 \\ \nabla_q L &= \nabla_q \phi(w,q) - \frac{\mu}{2} \nabla_q ||q||_2^2 = 0 \\ \frac{dL}{d\lambda} &= 1 - ||w||_2^2 \\ \frac{dL}{d\mu} &= 1 - ||q||_2^2 \end{aligned}$$

3.2.3 Calcul des gradients

$\phi(w,q) = w^{'} X^{'}_{k-1} DY_{k-1} q$ d'où $\nabla_q \phi(w,q) = (w^{'} X^{'}_{k-1} DY_{k-1})^{'} = Y^{'}_{k-1} DX_{k-1} w$.
Et $\phi(w,q) = q^{'} Y^{'}_{k-1} DX_{k-1} w$ d'où $\nabla_w \phi(w,q) = (q^{'} Y^{'}_{k-1} DX_{k-1})^{'} = X^{'}_{k-1} DY_{k-1} q$.
De plus $||w||_2^2 = w^{'} w$ et $||q||_2^2 = q^{'} q$ alors $\nabla_w ||w||_2^2 = 2w$ et $\nabla_q ||q||_2^2 = 2q$.
Alors le système devient :

$$\nabla_w L = X^{'}_{k-1} DY_{k-1} - \lambda w = 0 \quad p \quad \textit{équations} \tag{3.1}$$

$$\nabla_q L = Y^{'}_{k-1} DX_{k-1} w - \mu q = 0 \quad c \quad \textit{équations} \tag{3.2}$$

$$\frac{dL}{d\lambda} = 1 - ||w||_2^2 = 0 \qquad 1 \quad \textit{équations} \tag{3.3}$$

$$\frac{dL}{d\mu} = 1 - ||q||_2^2 = 0 \qquad 1 \quad \textit{équations} \tag{3.4}$$

On a donc un système de $p + c + 2$ équations à $p + c + 2$ inconnus.

3.2.4 Calcul de λ et μ

$w^{'} * (3.1) = w^{'} X^{'}_{k-1} DY_{k-1} q - \lambda w^{'} w = 0$.
Or par (3.3) $w^{'} w = 1$ d'où :

$$w^{'} X^{'}_{k-1} DY_{k-1} q = \lambda.$$

De même $q^{'} * (3.2) = q^{'} Y^{'}_{k-1} DX_{k-1} w - \mu q^{'} q = 0$.
Or par (3.4) $q^{'} q = 1$ d'où :

$$q^{'} Y^{'}_{k-1} DX_{k-1} w = \mu.$$

Alors : $\lambda = \mu$.
Donc à l'optimum, on sait que :

$$\lambda = \mu = \phi(w, q).$$

3.2.5 Calcul de w et q

On appelle (3.1) et (3.2) les formules de transitions car on peut calculer w et q l'un en fonction de l'autre alors qu'ils n'appartiennent pas au même espace.
$\mu * (3.1) + X'_{k-1} D Y_{k-1} \mu q - \lambda \mu w = 0$.
Or $\lambda = \mu$ et d'aprés (3.2) $\mu q = Y'_{k-1} D X_{k-1} w$.
D'où :

$$X'_{k-1} D Y_{k-1} Y'_{k-1} D X_{k-1} w = \lambda^2 w \ .$$

De même $\lambda * (3.2)$ et $(3.1) + \lambda = \mu$, alors :

$$Y'_{k-1} D X_{k-1} X'_{k-1} D Y_{k-1} q = \lambda^2 q.$$

Parmi toutes les solutions du système (c-à-d parmi les vecteurs propre et les valeurs propres), je choisis la valeur propre la plus grande λ^2 pour les 2 matrices précédentes.

3.2.6 Propriétés des composantes $t_1, ..., t_K$

Les formules d'actualisation des variables conduisent à la relation

$$(t_k, t_l)_D = (t_k, u_l)_D = 0, \quad \forall l > k.$$

La non corrélation ou D-orthogonalité, mutuelle entre les composantes $t_1, ..., t_K$ a de multiples conséquences. On montre ainsi par récurrence que t_k appartient à ImX espace vectoriel engendré par les prédicteurs. Plus précisement, $t_k = X\alpha_k$ avec :

$$\begin{aligned} \alpha_1 &= w_1 \\ \alpha_k &= \left[\mathbf{I}_p - \sum_{j=1}^{k-1} \frac{\alpha_j \alpha'_j}{||t_j||^2_D} X' D X \right] w_k \quad \forall k > l. \end{aligned}$$

La non corrélation implique en outre que $\sum_{k=1}^{K-1} P_{t_k} = P_{T_K}$,

où $P_{T_K} = T_K(T_K' D T_K)^{-1} T_K' D$ est le projecteur orthogonal sur la matrice $T_K = [||t_1||, ..., ||t_K||]$.

Les deux derniers résultats permettent de considérer P_{T_K} comme le projecteur sur le sous espace de ImX engendré par les composantes $t_1, ..., t_K$. Dans le cas particulier où $K = rang(X)$, $P_{T_K} = P_X$.

3.2.7 Modèle PLS

Les formules d'actualisation entranent l'écriture des modèles linéaires :

$$\begin{aligned} X &= E_0 = \hat{X}_K + E_K \\ Y &= F_0 = \hat{Y}_K + F_K, \end{aligned}$$

où $\hat{X}_k = P_{t_k} E_{k-1}$ et $\hat{Y}_k = P_{t_k} F_{k-1}$ sont les modèles partiels de rang 1. $\hat{X}_K$ est l'approximation de X avec une erreur E_K, idem pour $\hat{Y}_K$.

L'actualisation des variables et la non corrélation des composantes conduisent à écrire plus simplement les modèles partiels : $\hat{X}_k = P_{t_k} X$ et $\hat{Y}_k = P_{t_k} Y$. La non corrélation des composantes conduit à l'écriture définitive des modèles PLS en fonction des composantes :

$$\begin{aligned} \hat{Y}_K &= P_{T_K} Y. \\ \hat{X}_K &= P_{T_K} X. \end{aligned}$$

Le projecteur s'écrit aussi :

$$P_{T_K} = \sum_{k=1}^{K} \frac{X \alpha_k \alpha_k' X' D}{||t_k||_D^2},$$

ce qui implique que le modèle PLS est linéaire en les variables explicatives initiales :

$$\hat{Y}_K = X \hat{\beta}_K,$$

avec :

$$\hat{\beta}_K = \sum_{k=1}^{K} \frac{\alpha_k \alpha_k^{'}}{||t_k||_D^2} X' DY.$$

3.2.8 Recherche de la taille K

On recherche une taille de modèle K, ou ici un nombre de composantes K, qui soit compris entre 1 et une taille maximum A. Cette taille maximum peut être choisie comme $A = rang(X)$ ou comme la taille au-delà de laquelle il est certain que les composantes ne serviront à rien.

Méthode graphique

Une première méthode consiste à tracer un diagramme d'évolution des coefficients $\beta^*_{(j)}$ en fonction du nombre j de composantes. Cette méthode visuelle possède l'inconvénient majeur de n'avoir aucun support analytique d'aide à la décision.

Apprentissage-validation.

La procédure d'apprentissage-validation consiste à séparer de manière aléatoire les données en deux parties distinctes (X_a, Y_a) et (X_v, Y_v). Une régression PLS est conduite avec le jeu d'apprentissage (X_a, Y_a) pour toutes les tailles de modèles possibles. Ensuite, en utilisant tous ces modèles et les variables explicatives X_v, les valeurs de la variable à expliquer sont prédites $\hat{Y}_v(j)$ pour toutes les tailles j. La qualité du modèle est ensuite obtenue en mesurant la distance entre les observations prévues et les vraies observations par un critère. Le plus connu est le PRESS :

$$PRESS(j) = ||\hat{Y}_v^{PLS}(j) - Y_v||^2.$$

D'autres critères peuvent être utilisés comme :

$$MAE = ||\hat{Y}_v^{PLS}(j) - Y_v||_1.$$

La taille optimale K choisie est celle qui conduit à la minimisation du critère choisi.

Conclusion

Les composantes PLS $t_1, t_2, ..., t_K$ sont donc des combinaisons linéaires des colonnes de X (matrice centrée-réduite des variables initiales), non corrélées entre elles, résumant au mieux X tout en expliquant autant que possible Y (vecteur centré-réduit de la variable réponse initiale).

Ces composantes sont donc analogues à des composantes principales des $X_1, X_2, ..., X_p$ (les p variables explicatives initiales) expliquant au mieux la variable réponse initiale.

Chapitre 4

Simulation

4.1 Exemple des biscuits

Cet exemple est cité par Brown et al. (2001)[7] et les données sont disponibles sur la page personnelle de M. Vannucci (www.stat.tamu.edu/ mvannucci/) et sur le package **fds** du logiciel R.

4.1.1 Données

Nous sommes en présence de biscuits non cuits pour lesquels on souhaite connaître rapidement et à moindre coût, la composition en quatre ingrédients : les lipides, les sucres, la farine et l'eau. Des méthodes classiques de chimie analytique permettent de mesurer la composition des biscuits mais elles sont assez longues et coûteuses et ne peuvent pas être mises en ligne sur une chaîne de production. Il serait souhaitable de pouvoir les remplacer par la mesure d'un spectre d'absorbance dans le domaine proche infrarouge (ou spectre proche infrarouge). Pour savoir si cela est possible, nous allons devoir essayer d'expliquer la composition par le spectre.

Nous avons $n_a = 40$ biscuits non cuits sur lesquels sont mesurés les spectres proches infrarouges : on mesure l'absorbance à une longueur d'onde donnée, pour toutes les longueurs d'ondes entre 1100 et 2498 nanomètres et régulièrement espacées de 2 nanomètres. Nous avons donc 700 variables potentiellement explicatives. Ensuite, pour chaque biscuit, on mesure sa composition par les méthodes traditionnelles. Ici nous

allons nous intéresser uniquement au pourcentage de sucres. Nous avons donc $p = 700$ variables pour $n_a = 40$ individus.

Comme nous souhaitons savoir si l'on peut vraiment expliquer le taux de sucres par le spectre proche infrarouge, nous disposons d'un échantillon de validation pour comparer les méthodes. Cet échantillon de validation comporte $n_v = 32$ individus et ne sera jamais utilisé pour estimer les coefficients d'un modèle quel qu'il soit. Il sert uniquement à comparer une méthode avec une autre et à connaître, pour une méthode, sa capacité de prévision. Cette séparation en deux échantillons de tailles 40 et 32 fait partie du jeu de données et nous ne nous poserons donc pas la question de cette répartition.

Les ordres permettant d'importer les données sont les suivants :

```
> library(fds)
Loading required package: rainbow Loading required package: MASS
Loading required package: pcaPP Loading required package: RCurl
Loading required package: bitops
> data(labp)
> data(labc)
> data(nirp)
> data(nirc)
>
> Xbrut.app <- t(nirc$y) La matrice des variables explicatives,
échantillon d'apprentissage.
> colnames(Xbrut.app)=paste("spect",nirc$x,sep="")
> Ybrut.app <- t(labc) La matrices des variables réponses: les
variables mesurant le sucre dans les biscruits.
> colnames(Ybrut.app)=rownames(labc)
> Xbrut.val <- t(nirp$y)
> colnames(Xbrut.val)=paste("spect",nirp$x,sep="")
> Ybrut.val <- t(labp)
> colnames(Ybrut.val)=rownames(labc)
>
```

```
> Yselec <- 2 Variable à expliquer
> cookie.app <- cbind.data.frame(Ybrut.app[,Yselec],Xbrut.app)
> colnames(cookie.app)[1] <- "sucres"
>
> cookie.val <- cbind.data.frame(Ybrut.val[,Yselec],Xbrut.val)
> colnames(cookie.val)[1] <- "sucres"
```

4.2 Traitement des données

4.2.1 Régression linéaire multiple

La régression linéaire multiple est obtenue grâce à la fonction suivante :

```
> modele.lm=lm(sucres∼.,data = cookie.app)
```

Nous pouvons ensuite évaluer la capacité de prévision par le MSEP sur notre jeu de validation :

```
> prediction.mco<-predict(modele.lm,newdata=cookie.val)
> eqmp.lm<-mean((cookie.val[,1]-prediction.mco)^ 2)
```

Nous en déduisons que le MSEP sur le jeu de données de validation vaut 134.5219.

4.2.2 Conclusion

L'erreur moyenne de prévision par la régression linéaire multiple est donc très grande. Cela nous laisse penser que la matrice $X^{'}X$ est *mal conditionnée*.

4.2.3 Mise en évidence de corrélations entre variables

Plusieurs moyens existent pour prouver la présence de corrélations entre les variables. Ici, nous allons nous contenter de deux tests simples que sont la construction de la matrice $X^{'}X$ et le calcul de son déterminant.

Déterminant de $X'X$:

```
> det(t(X)%*%X)
[1] 8.777817e-12
```

Remarque 4.2.1. *La présence de corrélations (confirmées par $det(X'X) \approx 0$ et les forts coefficients de corrélation entre les variables) explique les résultats* aberrants *de l'estimation par MC.*

4.3 Régression dans le cadre de données corrélées

4.3.1 Régression sur composantes principales

Afin d'utiliser la régression sur composantes principales, nous devons déterminer le nombre de composantes à retenir. Ce nombre k sera toujours déterminé par validation croisée sur 4 groupes de 10 observations. Rappelons la méthode proposée par le package **pls**. Nous contrôlons la graine du générateur afin d'obtenir toujours la même partition pour toutes les méthodes de ce chapitre.

```
> library(pls)
> set.seed(87)
> cvseg <- cvsegments(nrow(cookie.app),k=4,type="random")
```

La régression sur composantes principales est conduite simplement grâce à la fonction **pcr**. Ici nous pouvons avoir au maximum 40 composantes principales ($\min(n_a, p) = n_a = 40$), mais nous avons choisi un nombre maximum un peu moins grand ($K = 30$) pour des raisons de présentation graphique.

Afin d'utiliser les mêmes estimateurs de variance empiriques, calculons ceux-ci sur les variables explicatives.

```
> n.app <- nrow(cookie.app)
> stdX.app <- sqrt(apply(cookie.app[,-1],2,var)*(n.app-1)/n.app)
```

La modélisation est enfin obtenue grâce aux ordres ci-dessous :

```
> modele.pcr <- pcr(sucres~.,ncomp=30,data =
cookie.app,scale= + stdX.app,validation = "CV",segments=cvseg)
> msepcv.pcr <- MSEP(modele.pcr,estimate=c("train","CV"))
```

```
> x=c(msepcv.pcr$val[1,,],msepcv.pcr$val[2,,])
> y=c(rep("train",length(msepcv.pcr$val[1,,])),rep("cv",length(msepcv.pcr$val[2,,])))
> z=c(0:29,0:29)
> dt=data.frame(x,y,z)
> colnames(dt)=c("msep","sample","comps")
library(ggplot2)
> p<-ggplot(dt,aes(x=comps,y=msep,col=sample))+geomline()
> p+themebw()
> plot(explvar(modele.pcr),type="l",main="")
```

Cette fonction centre et réduit les variables et calcule aussi la MSEP pour la validation croisée. Nous faisons figurer aussi la part de variance des X prise en compte par chaque composante. Dès la i$^{\text{ième}}$ composante, la part de variance des X expliquée par chaque composante est quasi nulle. Il ne subsiste que peu de variabilité initiale non prise en compte dans le modèle. Le nombre de composantes k est trouvé numériquement par :

```
> ncomp.pcr <- which.min(msepcv.pcr$val["CV",,])-1
```

et vaut 6, valeur que nous retrouvons sur le gaphique suivant :

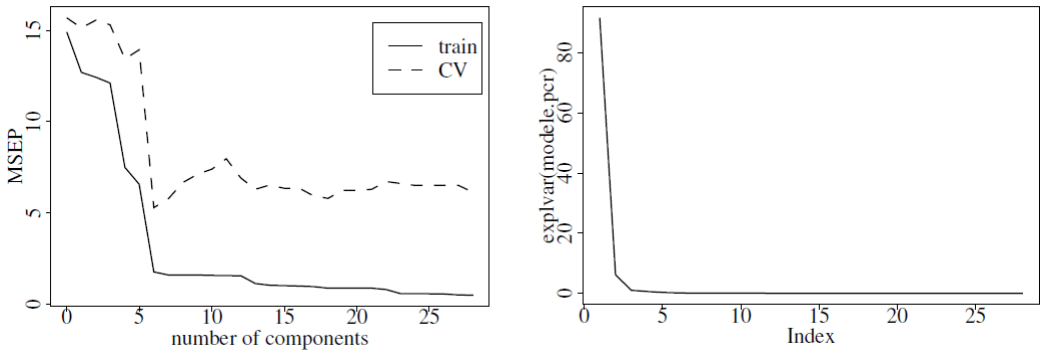

Figure (4.1) Evolution du MSEP en fonction du nombre de composantes de la régression sur composantes principales (graphique de droite). Evolution de la part de variance (en %) des X prise en compte par chaque composante.

Le graphique des résidus ne montre aucune structuration particulière et nous ne le reproduirons pas ici. La prévision par le modèle des observations du jeu de validation est obtenue par :

```
> modele.pcr.fin <- pcr(sucres∼.,ncomp=ncomp.pcr,data =
cookie.app, + scale=stdX.app)
> ychap <-predict(modele.pcr.fin,newdata=cookie.val)[,1,ncomp.pcr]
> res.pcr <-cookie.val[,"sucres"]-ychap
> mean(res.pcr ^ 2)
```

Nous en déduisons que le MSEP sur le jeu de données de validation vaut 1.03.

4.3.2 Régression PLS

Pour cette méthode de régression, et à l'image de la régression sur composantes principales, nous devons déterminer le nombre k de composantes PLS grâce aux ordres ci-dessous :

```
> modele.pls <- plsr(sucres∼.,ncomp=30,data = cookie.app,scale= +
stdX.app,validation = "CV",segments=cvseg)
```

Le vecteur `stdX.app` contient les écarts-types empiriques.

Le choix du nombre de composantes est réalisé graphiquement par :

```
> msepcv.pls <- MSEP(modele.pls,estimate=c("train","CV"))
> x=c(msepcv.pcr$val[1,,],msepcv.pcr$val[2,,])
> y=c(rep("train",length(msepcv.pcr$val[1,,])),rep("cv",length(msepcv.pcr$val[2,,])))
> z=c(0:29,0:29)
> dt=data.frame(x,y,z)
> colnames(dt)=c("msep","sample","comps")
library(ggplot2)
> p<-ggplot(dt,aes(x=comps,y=msep,col=sample))+geomline()
> p+themebw()
> plot(explvar(modele.pls),type="l",main="")
```

Ce minimum est obtenu simplement par :

```
> ncomp.pls <- which.min(msepcv.pls$val["CV",,])-1
```

La représentation graphique nous indique que 5 composantes pourraient donner un résultat presque aussi bon que le minimum numérique qui est de 10 composantes.

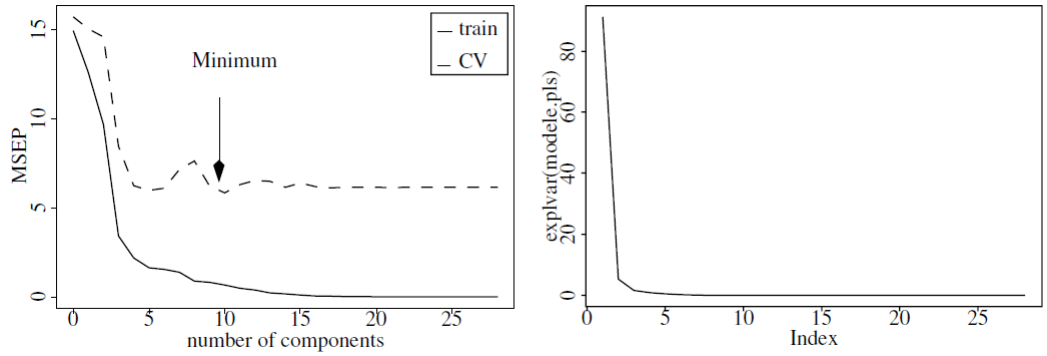

Figure (4.2) Evolution du MSEP en fonction du nombre de composantes de la régression sur composantes PLS (graphique de droite). Evolution de la part de variance (en %) des X prise en compte par chaque composante.

Nous pouvons ensuite évaluer la capacité de prévision par le MSEP sur notre jeu de validation :

```
> modele.s.fin <- plsr(sucres~.,ncomp=ncomp.pls,data=cookie.app, +
scale=stdX.app)
> ychap <- predict(modele.pls.fin,newdata=cookie.val)[,1,ncomp.pls]
> res.pls <- cookie.val[,"sucres"]-ychap
> mean(res.pcr ^ 2)
```

Cela donne un MSEP d'environ 4. Si le modèle parcimonieux à 5 composantes avait été choisi, alors le MSEP serait de 0.78 chiffre plus faible (et donc meilleur) que celui de 10 composantes. Cette remarque montre bien la difficulté de choisir le nombre de composantes.

4.3.3 Comparaison des méthodes à partir des résidus

MLR		PCR		PLS	
Y	$\hat{Y}$	Y	$\hat{Y}$	Y	$\hat{Y}$
16.44	13.4714568	16.44	16.54214	16.44	13.846361
13.49	16.1288904	13.49	13.22690	13.49	10.962243
20.24	2.3668581	20.24	19.45944	20.24	17.294921
19.82	0.8862277	19.82	19.83156	19.82	18.957823
17.29	15.7486720	17.29	17.34016	17.29	16.268631
21.93	18.6515900	21.93	21.33869	21.93	19.651922
16.02	16.0250403	16.02	15.29858	16.02	15.050375
12.65	21.1254497	12.65	13.29156	12.65	10.589016
18.13	8.9509438	18.13	16.73438	18.13	24.888914
20.66	-18.2068741	20.66	19.23718	20.66	23.109766
10.96	16.4825213	10.96	11.93553	10.96	10.407821
14.34	-2.3739973	14.34	12.65592	14.34	15.165265
11.81	23.3523145	11.81	10.63131	11.81	12.009319
12.23	5.4670770	12.23	11.99230	12.23	14.605651
15.18	10.2033573	15.18	13.42538	15.18	14.740740
22.77	13.0381039	22.77	21.14052	22.77	26.201255
10.12	7.6065560	10.12	9.40280	10.12	10.089006
13.91	22.6356099	13.91	13.17462	13.91	12.624565
22.35	19.7847799	22.35	22.59705	22.35	21.799665
21.08	10.6788214	21.08	21.46227	21.08	22.842448
21.50	3.6095411	21.50	21.88306	21.50	22.726314
17.71	26.2278018	17.71	16.58377	17.71	18.273774
11.38	8.7815847	11.38	12.31299	11.38	9.660274
14.76	30.6866301	14.76	13.45267	14.76	13.667430
15.60	23.9403382	15.60	14.06300	15.60	15.874957

Table 4.1- Tableau des Y observés et des prévisions $\hat{Y}$ associées à partir les 3 méthodes.

Résidus issus de la MLR	Résidus issus de la PCR	Résidus issus de la PLS
2.968543234	-0.10213502	0.11253444
-2.638890383	0.26310213	0.49953983
17.873141947	0.78056182	1.02070186
18.933772314	-0.01155944	0.18781850
1.541328026	-0.05015817	0.09881349
3.278410037	0.59130797	0.76896626
-0.005040273	0.72141777	0.70280182
-8.475449696	-0.64156066	-0.72860369
9.179056226	1.39562346	0.29012504
38.866874113	1.42282079	0.78152828
-5.522521275	-0.97552505	-1.32453753
16.713997260	1.68408267	1.17694030
-11.542314510	1.17868584	1.04352154
6.762923000	0.23769541	-0.27864239
4.976642656	1.75462241	1.39866995
9.731896121	1.62947772	0.87435137
2.513443994	0.71719983	0.32346535
-8.725609934	0.73538391	0.55087699
2.565220129	-0.24704963	-0.54733228
10.401178563	-0.38227085	-0.83961219
17.890458933	-0.38305624	-0.66685953
-8.517801771	1.12623182	0.76278113
2.598415272	-0.93299292	-0.82019352
-15.926630106	1.30732856	1.10799387
-8.340338173	1.53700398	1.13093727
0.655623257	0.78338900	0.60612586
-4.436106084	1.04230299	0.75718041
-2.197643289	0.22989575	-0.17273654

Table 4.2 - Comparaison des résidus des régressions.

MSEP MLR	MSEP PCR	MSEP PLS
134.5219	1.027865	0.7807496

Table 4.3 - Comparaison des qualités prédictives des modèles issus des régressions sur MLR, PCR et PLS.

Conclusion

Nous en déduisons que l'erreur moyenne de prévision (MSEP) de la régression aux moindres carrés partiels (PLS) est meilleure que la régression linéaire multiple ou la régression sur composantes principales.

Ainsi le MSEP pour le modèle MLR est 134.521903 et pour le modèle PLS est 0.7807496 ce qui montre que la régression PLS apport une amélioration considérable à la régression linéaire multiple.

La prévision par proche infrarouge du taux de sucres semble assez satisfaisante, à condition de bien choisir la méthode de régression.

4.4 Conclusion et perspectives

Dans ce travail on a vu le modèle de régression linéaire multiple et l'estimation des paramètres de ce modèle par la méthode des moindres carrés on a vu aussi que la multicolinéarité des variables s'impose un grand problème dans l'estimation par MC.

Comme solution de ce problème on a vu deux méthodes la première c'est la régression sur composantes principales (PCR), la pratique de la méthode est fait une ACP (Analyse sur Composantes Principales) d'ordre k du tableau des données X qui va donner k composantes principales qui sont des combinaisons linéaires des variables naturelles et ne sont pas corrélées ensuite on effectue la régression sur ses composantes.

La deuxième méthode c'est la régression aux moindres carrés partiels (PLS) et de même comme la régression sur composantes principales l'objectif de la régression PLS est de construire des nouvelles variables qui soient combinaison linéaire des variables initiales, de telle sorte que ces composantes sont les plus corrélées avec la variable réponse Y.

Enfin, il est important de noter que dans le cas d'existence d'une colinéarité entre les variables explicatives, les modèles de régression RCP et PLS assurent des qualités de prédiction du phénomène meilleures que le modèle de régression linéaire multiple.

Perspectives

- La régression PLS non linéaire NLPLS ([26])([28]).
- La régression SPLPLS ([30]).
- PLS et méthodes de Lanczos ([02]).
- PLS et gradients conjugués ([30]).
- L'algorithme PLS-Cox ([02]).
- Lasso Regression ([29]).

Bibliographie

[1] Angrist, J. et Pischke, J-S. (2008), Mostly Harmless Econometrics : An Empiricist's Companion, Princeton University Press.

[2] Bastien, P. (2008), Régression PLS et données censurées, PhD thesis, Conservatoirenational des arts et métiers de Paris.

[3] Bourbonnais, R. (1998), Économétrie, Collection Eco sup, Dunod.

[4] Bourbonnais, R. (2018), Économétrie, Dunod,10ème Edt.

[5] Cameron, C. et Trivedi, P. (2005), Microeconometrics : Methods And Applications, Cambridge University Press.

[6] Confais, J. et Le Guen, M. (2006), Premiers pas en régression linéaire, La Revue Modulad, N°35, pp 220-363.

[7] Cornillon, P-A. et Eric, M-L. (2007), Régression : Théorie et applications, Springer.

[8] Crépon, B. et Jacquemet, N. (2010), Économétrie : Méthode et Applications, De Boeck Université, coll. Ouvertures économiques.

[9] Dodge, Y. et Rousson, V. (1994), Analyse de régression appliquée, Dunod,(2004).

[10] Dodge, Y. (2010), The Concise Encyclopaedia of Statistics, New York, Springer.

[11] Dormont, B. (2007), Introduction à l'économétrie, Paris, Montchrestien.

[12] Foucart, T. (2006), Colinéarité et régression linéaire, Mathématiques et sciences humaines, vol 1, 5-25.

[13] Galton, F. (1886), Regression Towards Mediocrity in Hereditary Stature, *Journal of the Anthropological Institute*, vol 15, 246-263.

[14] Gelman, A. et Hill, J. (2006), Data Analysis Using Regression And Multilevel, Hierarchical Models, Cambridge University Press, coll. Analytical Methods for Social Research.

[15] Giraud, R. (1994), N. Chaix, Économétrie,Puf.

[16] Hastie, T. , Tibshirani, R. et Friedman, J. (2009), The Elements of Statistical Learning : Data Mining, Inference, and Prediction.

[17] James, G. (2013), Daniela Witten, Trevor Hastie et Robert Tibshirani, An Introduction to Statistical Learning, Springer Verlag, coll.

[18] Jong, S-D (1993), An alternative approach to partial least squares regression.Chemometrics and Intelligent Laboratory Systems.

[19] Kondylis, A. (2006). PLS methods in regression - Model assessment and inference. Phd thesis, Insitut de statistique - Faculté des sciences économiques - Université de Neuchâtel.

[20] Labrousse, C. (1972), Introduction à l'économétrie Maîtrise d'économétrie, Dunod, Paris.

[21] Legendre, A-M (1805), Nouvelles méthodes pour la détermination des orbites des comètes, Paris, F. Didot.

[22] Lingren, F. , Geladi, P. and Wold, S. (1993), The kernel algorithm for pls. *Journal of Chemometrics*, vol 7 , 45-59. Springer Texts in Statistics.

[23] Mignon, V. (2008), Économétrie, Economica, coll. Corpus économie.

[24] Palm, R. et Iemma, A-F.(1995), Quelques alternatives à la régression classique dans le cadre de la colinéarité, Revue de statistique appliquée, vol 43, 5-33.

[25] Rannar, S. , Lindgren, F. , Geladi, P. and Wold, S. (1994). A pls kernel algorithm for data sets with many variables and fewer objects. part i : Theory and algorithm. *Journal of Chemometrics*, vol 8 ,111-125.

[26] Rosipal, R. (2011), Nonlinear Partial Least Squares : An Overview.

[27] Saporta, G. (1990), Probabilités, analyse des données et statistique. Technip.

[28] Tenenhaus, M. (1998), La régression PLS : Théorie et Pratique, Paris.

[29] Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society*, vol 58, 267-288. éditions Technip.

[30] Vivien, M. (2002), Approches PLS linéaires et non linéaires pour la modélisation de multi-tableaux : théorie et applications.

[31] Wasserman, L. (2004), All of Statistics : A Concise Course in Statistical Inference, New York, Springer-Verlag.